DOCUMENT RESUME

ED 177 210 TM 009 717

AUTHOR Rosenbach, John H.
TITLE Instrumentation for Assessing Culturally Different: Some Conceptual Issues.
PUB DATE Apr 79
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Aptitude Tests; Cultural Differences; Culture Free Tests; Elementary Secondary Education; Higher Education; Intelligence Differences; *Intelligence Tests; *Predictive Validity; Racial Differences; *Relevance (Education); Social Differences; *Test Bias; *Testing Problems

ABSTRACT

Since their beginning, intelligence tests have favored the higher social classes. Despite federal mandates to the contrary, bias in assessment is likely to remain a problem. Claims of test bias can be categorized as popular (naive); clinical (intuitive); statistical (predictive); and psychometric (construct and content). A literature review has led the author to believe that most claims of test bias are unfounded, especially for intelligence and aptitude tests which are used to predict success in school. Since most of these tests are valid and remarkably accurate in their predictions, the bias must be associated with the nature of what is being predicted (the Criteria). Although tests are accurate predictors of school achievement, success in school is not very predictive of success outside of school. Alternative solutions to this issue are to explore aptitude-treatment-interaction models which may be used to improve the test performance of minority group students, or to modify our expectations of what should be learned in school. Before the problems of test bias can be resolved, the sources of bias must be objectively determined and the purposes of testing must be understood. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177210

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Instrumentation for Assessing Culturally Different: Some Conceptual Issues*

John H. Rosenbach

State University of New York at Albany

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John H. Rosenbach

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

From the time of Binet, Simon, and Stern it has been known that so-called intelligence tests favor the higher social classes. For example, Stern (1914) noted that the average test performance of lower class ten year olds equalled that of average nine year olds from higher classes. And despite repeated efforts to "equalize" test performance (see, example, the work of Eells, Davis, et al (1951) higher or middle class children continue to score, on the average, close to a standard deviation or so higher than those from the lower classes.** If the problem were one simply involving technical or psychometric considerations, it would have been resolved long ago, for developments of this kind have been awesome. But the problem has not been resolved and despite legislative mandates to the contrary, is likely to be with us for years to come, if not forever. It is the purpose of this paper to review some of the approaches to test (or assessment) bias and to direct attention to what this author

TM009 212

---

*Presented as part of a symposium, "Instrumentation for Assessing Culturally Different," NCME, April 9-11, 1979, San Francisco, California.

**In the following discussion it is assumed that tests are administered intelligently and appropriately (e.g., that one does not foolishly and stupidly give a test in the wrong language).

considers to be its more fundamental "causes." In fact, it is contended that our very fascination with technical (statistical and psychometric) advances has tended to obscure the roots of the problem and to have directed efforts away from its resolution.

Although the literature on test bias is multi-faceted it can be subsumed under the following (somewhat arbitrary) headings: popular (naive); clinical (intuitive); statistical (predictive); psychometric (construct); psychometric (content).

Popular (Naive). The fact that certain groups consistently score higher or lower than others on various aptitude tests is, in itself, sufficient evidence in the minds of many people to support a claim of bias. Usually they buttress this argument by pointing to the heavy verbal loading of the tests, suggesting that it is in the language area that lower scoring groups are most disadvantaged. This position, of course, ignores a rather substantial body of research that, if anything, suggests the opposite, i.e., mean differences are often even greater on tests that are less verbal and less obviously culturally loaded (e.g., Jensen, 1969).

Nonetheless, as simple, or as simple-minded, as this position is, it carries a great deal of force and, in part, is responsible for various court cases.

Clinical (Intuitive). In a somewhat different vein there are the observations of numerous clinicians who, in working with individual children, become convinced that many are more capable

than the results of a standardized test indicate; given more liberal time limits, or perhaps certain prompts or cues, or maybe even different questions or items, these children would demonstrate considerably higher ability. Mercer's work (1973), for example, seems to be an outgrowth of this point of view, and represents an interesting and potentially fruitful line of research. Regrettably, however, there is little evidence to support the usefulness of this approach; the subjects may appear to be more able on the basis of assessments but the test results have yet to be shown to correlate with educational criteria.

Statistical (Predictive). For at least twenty-five years the predictive-bias of tests has been examined with respect to whether a common regression line is appropriate for more than one group. First suggested by Humphreys (1952), this definition has become closely associated with Cleary (Cleary, et al., 1975). Mostly it has been used to compare advantaged and disadvantaged (usually whites vs blacks) groups in college settings. In general, the results of the research suggest that, for the purpose of predicting college success, aptitude tests (usually the SAT) are relatively unbiased and, if anything, tend to overpredict success for blacks, a finding which tends to be contrary to popular belief.

Although there are several alternative definitions of test bias, based on predictive models (Linn (1973)), none would seem to give strong support to the charge that most of the commonly used aptitude tests are unfair to minority groups or to lower SES classes.

Psychometric (Construct). An ingenious approach to test bias is presented by Jensen (1976), in an article entitled, "Test Bias and Construct Validity." Using such well-known instruments as the Peabody Picture Vocabulary Test, the Raven's Matrices, the Stanford-Binet, the WISC, and the Wonderlic Personnel Test, Jensen examined the internal structure of these tests for white and black groups. He concluded,

> The results are unequivocal: None of the several objective indices of cultural bias shows any significant indication of bias in any of these tests when they are used with blacks and whites. Correlations of raw scores with age, internal consistency reliability, rank order of item difficulty (i.e., percent passing), relative difficulty of adjacent items, item correlation with total score, loadings of items or tests on the general factor, and relative frequencies in choice of error distractors--all are substantially the same in the white and black groups. (Jensen, 1976, p. 345)

Psychometric (Content). One of the most ambitious efforts to develop a culture-fair intelligence test was mounted by the black sociologist, Allison Davis, in the 1940's and 1950's (Eells, K., Davis, Al, et al. (1951, Hess, (1955)). The final product, called the "Davis-Bells Games," was made up of a series of essentially non-verbal items selected on the basis of their commonality among social classes; the test allegedly measured problem-solving ability. But despite the careful attempt to use only items which were fair to the lower classes, subsequent "investigators found as much social-class difference on the Davis tests as on traditional ones....(and the) tests were not good predictors of success in school...." (Cronbach, 1975, p8). Nonetheless, Davis' pioneering work, especially his penetrating analysis of the cultural bias of schooling, is impressive; and it

is a sad, if not tragic, commentary on the state of education and psychology that his contributions have been overlooked by most contemporary writers. If Davis' work had appeared in the sixties, he would undoubtedly be a heroic figure today, but he had the misfortune of being concerned with issues that were ten to fifteen years ahead of the times; in the 40's and 50's very few influential people were speaking out over the plight of the poor, the blacks, and the disadvantaged.

* * * * * *

The point I have been trying to lead up to is that most claims of "test bias" are basically unfounded, especially for those tests which are used to predict success in school. The fact is most aptitude or intelligence tests are remarkably accurate in their predictions; to attack the tests because of what they foretell is simply reminiscent of the emperor who executed bearers of bad news. If there is bias, and *'m convinced that there is, then it must be in the nature of what is being predicted, i.e., in the criteria.

At this point it is important to remind ourselves that schooling, as we know it in America, is the product of the affluent, of the middle-and-higher social classes. It is also important to inquire as to the nature of schooling, i.e., the kinds of tasks children are expected to perform and the rationale for their inclusion in the curriculum. Unquestionably, certain of these tasks, e.g., those that deal with skills of basic reading and arithmetic, represent abilities that are required for survival or

near-survival in a modern society, but the remainder of the school curriculum is composed largely of tasks (learnings) that reflect a particular set of values, and often are justified on the grounds that they are "good" in and of themselves (e.g., a liberal education). The gravity of this situation is heightened when one is faced with the sobering fact that research has failed to demonstrate that success in school is related to anything else of significance (see, e.g., McClelland (1973) or Wallach & Wing (1969)). What we have, then, is a situation where tests are accurate predictors of school success (grades, scores on achievement tests) but where school success, in turn, is not very predictive of success outside of school ! Just such an analysis led McClelland to conclude, "....(it is) abundantly clear that the testing movement is in grave danger of perpetuating a mythological meritocracy in which none of the measures of merit bears significant demonstrable validity with respect to any measures outside of the charmed circle" (1973, p 2).

This is, of course, especially harsh criticism and may very well overstate the case against the relevance of schooling. But even granting that the research on this issue may have failed to detect certain long-range influences of school learning, it is probably not inaccurate to say that, within the higher social classes, the importance of education, as it is constituted in American schools, has been exaggerated. Without question an enormous amount of time in school is devoted to tasks that have little bearing on what happens outside the school. And these tasks possess certain common

features; they are school-like and usually demand something akin to "g" or general intelligence, i.e., they are abstract, symbolic, and complex. In essence, they reflect characteristics that are highly valued in their own right in the "educated" middle and upper classes. Some writers (e.g., McClelland (1973)) have even gone so far as to characterize schooling as a set of academic "games," with rules governing their play which differ little in nature from those used in such activities as chess, bridge, football or basketball, and in which winning or losing is determined largely by abilities or talents (genetically and/or environmentally produced) that the player brings to the game. And it is significant to note that even though coaches are prone to rationalize the importance of various sports, e.g., they build character or self-discipline, games are, for the most part, played for their own sake, i.e., they need not be justified with respect to other life activities.

But regardless of whether or not one accepts the premise that schooling is mainly a game, the problem of criterion-bias still remains. And because the nature of school curricula is rooted deeply in strongly held ideologies, any efforts to cope with this issue are likely to generate social-political controversy. As Tyler (Eells, Davis, et al (1951)) has pointed out we have two alternatives--to accept the schools as they exist or to reject what exists and search for alternate learning experiences. If we continue to adopt the former, non-controversial approach, then improvements in the school performance of lower class children will have to rest on the hope that either some forms of early intervention

programs will make lower-class or minority children more like middle-class whites or something in the nature of ATI models can be developed which exploit the learning and performance abilities of these children. If we were to be successful in raising both IQ and achievement, nothing especially would have to be done about the use of existing aptitude tests; validity coefficients would be unchanged and the grounds for claims of test bias would disappear. If ATI models proved successful, however, there would be interesting implications with respect to validity of current tests, because the correlations between predictors and outcomes would necessarily fall; to maintain high validity, either moderator variables would have to be used or new tests would have to be constructed.

The second approach, is of course, loaded with controversy for it would require a major shift in education objectives, either through replacing existing curricula or in offering curriculum options which currently do not exist. There is a bit of irony in the fact that Allison Davis' analysis of schooling led him to a conclusion that is somewhat similar to that of Arthur Jensen's, i.e., that schooling should be redesigned to be less verbal and less middle-class culture-laden so as to allow for lower-class children to use their learning and reasoning abilities. Jensen, of course, did not extend his argument to include modifying curricula of middle-class children nor does he speak of capitalizing on the reasoning abilities of the lower classes. But each has suggested that we modify our expectations as to <u>what</u> is to be learned in school.

* * * * * *

A first step towards resolving problems of test bias must be to attempt to understand the sources of bias through careful, objective study of what is to be assessed and the purposes underlying the assessment. Emotionally based reactions to the problem, however well intended or as politically attractive as they may be, can only result in confusion and in creating false hopes.

With respect to the educational sphere, we have seen that the standard measures of ability, aptitude, or intelligence, be they individually or group administered, are about as "valid" as the reliabilities of the predictors and criteria permit. But these measures share important characteristics: they are constructed to reflect two basic assumptions (i.e., the underlying variable is continuous and is distributed normally within the populations to be tested) and they tap similar, if not identical, intellectual processes (Rosenbach, 1978). Kelley (1927), it will be recalled, referred to this overlap (perhaps as much as 90 percent) as reflecting the "Jangle Fallacy." It also suggests, however, that one source of bias is within the achievement (criterion) measure, itself. In an earlier paper I argued that achievement tests are relatively poor indicators of learning, per se; rather, they are static measures of performance level and are not sensitive to change or gains in performance, which are central to any definitions of learning (Rosenbach, 1975). Thus, one approach to reducing bias is

through constructing tests that are designed to measure learning rather than achievement.

A second purpose for in-school testing is diagnostic, and it is less likely that issues of bias will raise in this use of tests. If carried out properly diagnosis should involve some form of task analysis and any instruments should derive inherently from what is to be learned. The problems encountered here are not so much psychometric as they are what to do once a diagnosis has been made. Our shortcomings are that we do not possess enough of an understanding of learning and instruction to take advantage of diagnostic information.

A third purpose of testing is certification, i.e., acknowledging to others that an individual(s) has demonstrated some level of acceptable performance that is allegedly significant to success in a subsequent setting. Two sources of bias are potentially present in this use of tests. First, the test itself, usually of an achievement type, is susceptible to bias because of what it does, as well as doesn't, measure (as discussed above), and second, unlike predicting educational achievement, efforts to predict out-of-school performance (e.g., job success) generally produce notoriously low validity coefficients. The question, however, of whether the certifying tests are biased is, perhaps, beside the point; in this instance the issue is criterion relevancy, i.e., does the test relate to components that are important to success? If the validity coefficients are low, as they usually are, then

the answer must be, "no"! And the solution is obvious: tests must be designed that are more sensitive to the criteria.

Conclusions

Significant advances in solving problems of test-bias or non-discriminatory assessment will not occur until we begin to focus on the nature of the criteria against which tests are validated. As long as tests are demonstrated to have high validity, any bias must derive from the criteria. In this case the problem is not one of psychological or psychometric origin and is not likely to yield to psychological or empirical attacks. The cultural bias of schooling, for example, is philosophically based and, ultimately, can be resolved only through social-political actions. Even in the case of achievement tests, which are not good measures of learning, the issue is whether society wishes to recognize learning or to continue to focus almost totally on end-performance. If measures of learning could be devised and used it is likely that many disadvantaged/handicapped students would show significant gains associated with school experience. And this would also reduce the predictive validity of existing tests. But, perhaps, changing the nature of what we measure in criterion behavior is less controversial than changing the content of school curricula.

In summary, if most of our widely used tests are biased it is because of biases in our culture. Modifying existing instruments or creating new ones, simply to produce similar score distributions across various sub-populations, is an

exercise in futility. Tests are useful only to the degree they contribute independent information in a decision making process, and this demands appropriately high levels of validity. This places professionals, such as school psychologists, who must contribute to decision making in cultural institutions in an uneviable position; as agents of schools and society they are under strong pressures, and even have certain obligations, to work within existing structures and to assist in maintaining the status quo. On the other hand, as professional psychologists and educators, they have a responsibility to perform their jobs according to sound research and theory. This is a dilemma that, at the moment, only the individual practitioner can resolve.

Bibliography

Cleary, T.A., Humphreys, L., Kendricks, S.A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

Cronbach, L.J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Eells, K., Davis, A., Havighurst, R., Herrick, V., & Tyler, R. Intelligence and cultural differences. Chicago: University of Chicago Press, 1951.

Hess, R.D. Controlling culture influence in mental testing: An experimental test. Journal of Educational Research, 1958, 49, 53-58.

Humphreys, L.G. Individual differences. In Annual Review of Psychology, Palo Alto: Annual Reviews, Inc., 1952, 3, 131-150.

Jensen, A.R. How much can we boost IQ and scholastic achievement. In Environment, Heredity and Intelligence, Harvard Educational Review, 1969, 1-123.

Jensen, A. R. A theory of primary and secondary familial mental retardation. In R. Ellis (Ed.), International Review in Mental Retardation. New York: Academic Press, 1970, 4, 33-105.

Jensen, A. R. Test bias and construct validity. Phi Delta Kappan, 1976, 58, 340-346.

Kelley, T. R. Interpretation of Educational Measurements. N.Y.: World Book Co., 1927.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

McClelland, D.C. Testing for competence rather than for intelligence. American Psychologist, 1973, 28, 1-14.

Mercer, J. P. Labeling the Mentally Retarded. Berkley: University of California Press, 1973.

Rosenbach, J. H. Instructional effectiveness: Some issues of teachability and learnability. New England Research Organization, Provincetown, Mass., May 1975.

Rosenbach, J. H. Sequential tests of educational progress, Series II, in O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook, Highland Park, N.J.: The Gryphon Press, 1978, pp. 91-95.

Wallach, M. & Wing, C. The talented student: A validation of the creativity-intelligence distinction. New York: Holt, Rinehart & Winston, 1969.